Korpuslinguistische Fallstudien zum Südtiroler Standardschriftdeutsch – das Projekt "Korpus Südtirol"

Stefanie Anstein (Bozen)

Abstract

In this paper the linguistically annotated "Korpus Südtirol" is presented, which can be used as a basis for systematic comparative studies of language varieties. Corpus linguistic methods and results as well as insights concerning the semi-automatic collection of special properties of the South Tyrolean German language – both on the lexical and on the phrasal level – are described. In these studies, automatically produced lists of objective differences provide a basis for further manual investigation and classification.

1 Hintergrund und Ziele des Beitrags

In der Autonomen Provinz Bozen-Südtirol in Norditalien sind ca. zwei Drittel der rund 500.000 Einwohner deutsche Muttersprachler, der überwiegende Teil davon Dialektsprecher. Italienische wie auch andere Muttersprachler, die in Südtirol Deutsch lernen, sind damit oft einem Sprachinput ausgesetzt, das sich von den anderen deutschen Varietäten (cf. Ammon 1995) unterscheidet.

Die Sprachensituation in Südtirol, wo nach dem Autonomiestatut[1] Italienisch und Deutsch als Amtssprache gelten, hat zur Folge, dass nahezu alle öffentlichen Bekanntmachungen zweisprachig verfasst sind (cf. Abb. 1). Sowohl bei der Kookkurrenz *Parkschein einholen* als auch etwa bei dem Beispiel *acquistare il biglietto a terra – Fahrschein nur im Vorverkauf* werden Abweichungen zum österreichischen oder bundesdeutschen Standard deutlich.[2]

Abb. 1: Zweisprachiges Hinweisschild in Bozen

[1] http://www.provincia.bz.it/aprov/suedtirol/autonomiestatut01_d.htm, Stand 20. November 2007.

[2] Auf der Webseite http://www.korpus-suedtirol.it sind weitere Beispiele, nähere Informationen zur Sprachensituation sowie weiterführende Literaturangaben zu finden.

Beispiele der gesprochenen Sprache sind z. B. der Ausdruck *sich* [z. B.] *Mayr schreiben*, um den Nachnamen zu nennen, oder die so genannte *m*-/*n*-Abrundung (cf. Schwienbacher 1997: 95) bei Dativen: *Interesse an den Service*.

Um zusätzlich zu bisherigen Studien (z. B. Rizzo-Bauer 1962, Riedmann 1972, Lanthaler/Saxalber 1995, Ammon et al. 2004, Abfalterer 2005) weitere Unterschiede des Südtiroler Deutschen zu anderen deutschen Varietäten systematisch zu erfassen, was bisher v. a. auf lexikalischer Ebene und teilweise in Einzelstudien geschah, können nun computerlinguistische Methoden und Werkzeuge besonders für die Untersuchung geschriebener Sprache angewandt werden. Im vorliegenden Beitrag wird dazu das Korpus Südtirol vorgestellt, das u. a. als Grundlage für datenbasierte systematische Studien dienen soll. Quantitative Analysen des Einflusses von Dialekten und Kontaktsprachen bieten die Möglichkeit, durchgeführte Studien zu überprüfen, zu untermauern und ggf. zu ergänzen. Im ersten Schritt kann dazu eine rein frequenzbasierte und statistische Auswertung der Unterschiede durchgeführt werden, die im nächsten Schritt manuell interpretiert und auf Ursachen hin untersucht werden.

2 Das Projekt "Korpus Südtirol"

Der Grundstein zum Korpus Südtirol wird derzeit in einer rund zweijährigen Pilotstudie "Korpusbasierte Sprachbeobachtung und Beratung: Untersuchungen zur deutschen Sprache in Südtirol vor dem Hintergrund sprachlicher Entwicklungstendenzen im deutschen Sprachraum" gelegt. Die Leitung dieses im Rahmen des Programms Interreg III-A Italien/Österreich finanzierten Projekts[3] liegt bei Prof. Johann Drumbl von der Freien Universität Bozen; weiterer Partner neben der Europäischen Akademie Bozen ist die Universität Innsbruck. Auf der Webseite http://www.korpus-suedtirol.it befinden sich weitere Hintergrund- und Projektinformationen sowie ein Link zur Suche in den bisher erstellten Korpora nach einer Registrierung. Auch weitere Partner sind aufgeführt, z. B. die Teilprojekte einer Initiative zur Erstellung des "C4"-Korpus, bestehend aus den vier Teilkorpora der jeweiligen Varietäten des Deutschen, entwickelt in Deutschland[4], Österreich[5], der Schweiz[6] und Südtirol[7].

Der Aufbau und die Erweiterung des Korpus Südtirol wurde und wird in folgenden Schritten durchgeführt. Nach der Formulierung der Forschungsfrage wurde die Textauswahl getroffen und die Urheberrechte mit den Verlagen geklärt. Falls keine elektronische Version eines Textes vorliegt, wird eine Digitalisierung durchgeführt, d. h. die Druckwerke werden eingescannt und mit einer speziellen Software zur Zeichenerkennung in elektronische Formate umgewandelt. Die elektronischen Daten werden in standardisierte XML-Dateien konvertiert und nach TEI-Richtlinien[8] strukturell und inhaltlich annotiert. Im Zuge einer systematischen Bestandsaufnahme der aufgenommenen und aufzunehmenden Texte wird eine umfassende Metadatenerhebung z. B. für die Subkorpuserstellung (Unterteilung in Texte aus bestimmten Textsorten, Dekaden, von bestimmten Autoren etc.) durchgeführt. Im nächsten Aufbereitungsschritt werden Absätze und Sätze als textstrukturelle Informationen und

[3] Unterstützt durch das Amt für Schulfinanzierung Bozen.

[4] Projekt DWDS - Das Digitale Wörterbuch der deutschen Sprache des 20. Jh. der Berlin-Brandenburgischen Akadamie der Wissenschaften: http://www.dwds.de.

[5] Projekt AAC – Austrian Academy Corpus der Österreichischen Akademie der Wissenschaften: http://www.aac.ac.at.

[6] Projekt SCHWEIZER TEXT KORPUS des Deutschen Seminars der Universität Basel: http://www.schweizer-textkorpus.ch.

[7] Projekt Korpus Südtirol: http://www.korpus-suedtirol.it.

[8] Text Encoding Initiative – Guidelines for Electronic Text Encoding and Interchange; http://www.teic.org/cms/index.xml.

Wortarten sowie Lemmata als linguistische Annotation hinzugefügt. Letztere Annotierung wird mit dem TreeTagger (Schmid 1994) vorgenommen. Um die so aufbereiteten Daten nach linguistischen Kriterien durchsuchbar zu machen, werden sie im letzten Schritt in eine Abfrageumgebung integriert. Bei den ersten Untersuchungen wurde die "Sketch Engine" (Kilgarriff et al. 2004) dafür verwendet; derzeit wird eine eigens angepasste Benutzeroberfläche[9] basierend auf dem System "Corpus Query Processor" (Evert 2005) entwickelt.

Ein Beispiel für die Suchanfragemaske der Sketch Engine ist in Abb. 2 dargestellt.

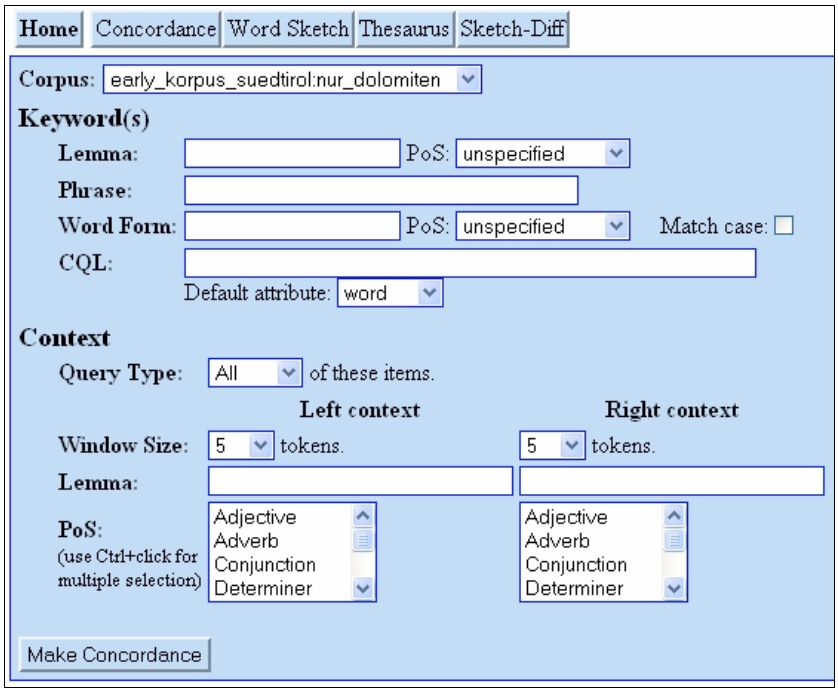


Abb. 2: Suchanfragemaske der Sketch Engine

Das Korpus Südtirol besteht bisher aus rund 68 Millionen Tokens, die vor allem aus Tageszeitungen und einigen Südtiroler belletristischen Werken stammen. Angestrebt ist eine gleichmäßige Verteilung auf die vier Textsorten Belletristik, journalistische Prosa, Gebrauchstexte und Fachliteratur, um die Südtiroler Varietät möglichst ausgeglichen und repräsentativ zu erfassen und ein so genanntes Referenzkorpus zu erstellen. Die Textsorten und weiteren Auswahlkriterien sind im Rahmen der oben genannten Korpusinitiative "C4" an die des Korpusprojekts DWDS angelehnt. Die Erfassung des geschriebenen Deutschen in Südtirol hat zudem eine kulturgeschichtlich relevante Archivierungsfunktion.

3 Methoden der Korpusuntersuchung und erste Ergebnisse

Die Textsammlung des in Kapitel 2 beschriebenen Korpus Südtirol kann nun z. B. für Vergleichsanalysen verwendet werden, um Besonderheiten und Unterschiede des Südtiroler Deutschen im Vergleich zu anderen Varietäten und somit in bestimmten Fällen eventuelle fehleranfällige linguistische Einheiten, wie z. B. den Gebrauch der Akkusativ- statt der Dativendung etwa im Ausdruck *Interesse an den Service*, zu identifizieren (zur Frage nach "Standard" und "Norm" cf. Riehl 1994). Dazu werden Textkorpora der verschiedenen Varietäten gegenübergestellt und parallel analysiert, um die Ergebnisse direkt miteinander vergleichen zu können. Spezifischen Unsicherheiten von Deutschlernenden in Südtirol (cf.

[9] Zugriff über http://www.korpus-suedtirol.it/index.php?id=14

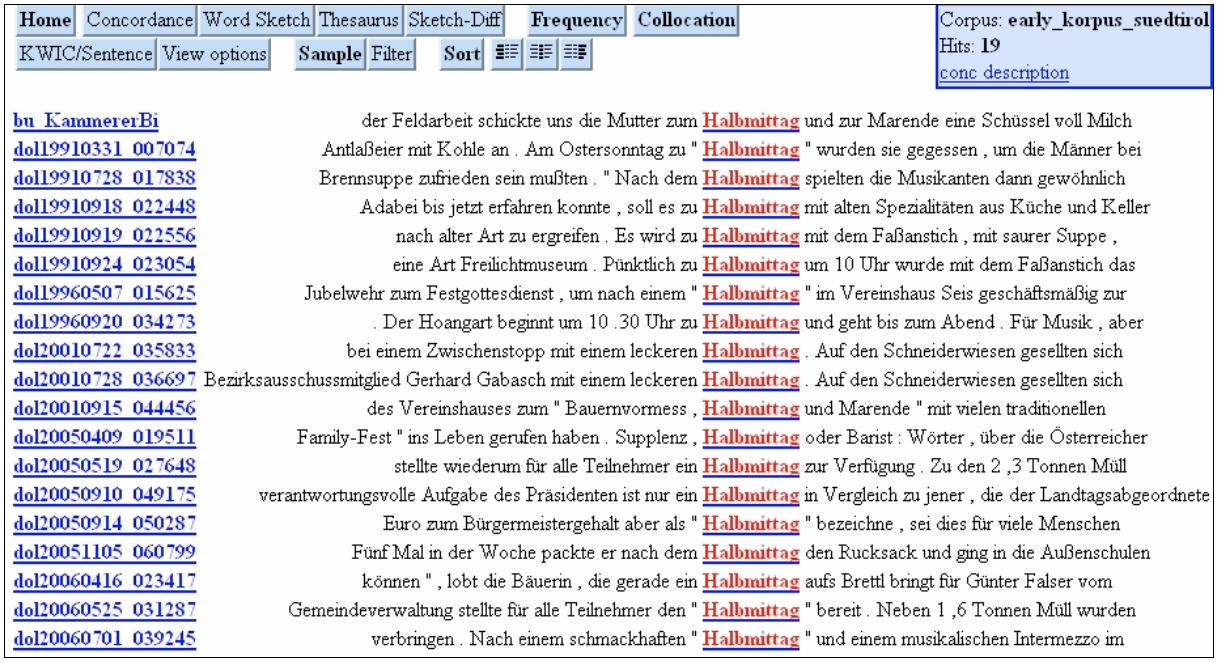
Saxalber Tetter 1989b) kann so mit Hilfe einer angepassten Sprachdidaktik gezielt entgegengewirkt werden; des Weiteren kann so auch das allgemeine Sprachbewusstsein der deutschen Muttersprachler in Südtirol gefördert werden.

Als Grundlage für die ersten exemplarischen Studien diente das rund 66 Millionen Tokens umfassende "Dolomiten-Subkorpus", das die Jahrgänge 1991, 1996, 2001, 2005 und 2006 der deutschsprachigen Südtiroler Tageszeitung "Dolomiten" enthält.

Im Folgenden werden die Ansätze zur Datengewinnung beschrieben und erste Ergebnisse der Korpusanalysen vorgestellt. Neben der Frage, wo genau Besonderheiten des Südtiroler Deutschen auftreten, wird auch kurz auf die Zuweisung von jeweiligen möglichen Ursachen eingegangen.

Die betroffenen linguistischen Ebenen, die hier beispielhaft semi-automatisch analysiert werden, sind (i) die Wortebene und (ii) die phrasale Ebene. Besonderheiten auf der phonologischen Ebene können mit einem geschriebenen Korpus offensichtlich nicht untersucht werden und für die semantische oder pragmatische Ebene sind ausgefeiltere computerlinguistische Methoden und mehr manuelle Vor- und Nachbearbeitung notwendig.

Auf der Wortebene unterscheiden wir lexikalische (Beispiel *Halbmittag*, cf. Abb. 3) und morphologische Phänomene (Beispiel *Geschenksidee*).



Home | Concordance | Word Sketch | Thesaurus | Sketch-Diff | Frequency | Collocation
KWIC/Sentence | View options | Sample | Filter | Sort

Corpus: **early_korpus_suedtirol**
Hits: **19**
conc description

bu_KammererBi	der Feldarbeit schickte uns die Mutter zum Halbmittag und zur Marende eine Schüssel voll Milch
dol19910331_007074	Antlaßeier mit Kohle an . Am Ostersonntag zu " Halbmittag " wurden sie gegessen , um die Männer bei
dol19910728_017838	Brennsuppe zufrieden sein mußten . " Nach dem Halbmittag spielten die Musikanten dann gewöhnlich
dol19910918_022448	Adabei bis jetzt erfahren konnte , soll es zu Halbmittag mit alten Spezialitäten aus Küche und Keller
dol19910919_022556	nach alter Art zu ergreifen . Es wird zu Halbmittag mit dem Faßanstich , mit saurer Suppe ,
dol19910924_023054	eine Art Freilichtmuseum . Pünktlich zu Halbmittag um 10 Uhr wurde mit dem Faßanstich das
dol19960507_015625	Jubelwehr zum Festgottesdienst , um nach einem " Halbmittag " im Vereinshaus Seis geschäftsmäßig zur
dol19960920_034273	. Der Hoangart beginnt um 10 .30 Uhr zu Halbmittag und geht bis zum Abend . Für Musik , aber
dol20010722_035833	bei einem Zwischenstopp mit einem leckeren Halbmittag . Auf den Schneiderwiesen gesellten sich
dol20010728_036697	Bezirksausschussmitglied Gerhard Gabasch mit einem leckeren Halbmittag . Auf den Schneiderwiesen gesellten sich
dol20010915_044456	des Vereinshauses zum " Bauernvormess , Halbmittag und Marende " mit vielen traditionellen
dol20050409_019511	Family-Fest " ins Leben gerufen haben . Supplenz , Halbmittag oder Barist : Wörter , über die Österreicher
dol20050519_027648	stellte wiederum für alle Teilnehmer ein Halbmittag zur Verfügung . Zu den 2 ,3 Tonnen Müll
dol20050910_049175	verantwortungsvolle Aufgabe des Präsidenten ist nur ein Halbmittag in Vergleich zu jener , die der Landtagsabgeordnete
dol20050914_050287	Euro zum Bürgermeistergehalt aber als " Halbmittag " bezeichne , sei dies für viele Menschen
dol20051105_060799	Fünf Mal in der Woche packte er nach dem Halbmittag den Rucksack und ging in die Außenschulen
dol20060416_023417	können " , lobt die Bäuerin , die gerade ein Halbmittag aufs Brettl bringt für Günter Falser vom
dol20060525_031287	Gemeindeverwaltung stellte für alle Teilnehmer den " Halbmittag " bereit . Neben 1 ,6 Tonnen Müll wurden
dol20060701_039245	verbringen . Nach einem schmackhaften " Halbmittag " und einem musikalischen Intermezzo im

Abb. 3: Anfrageergebnis für *Halbmittag*

Auf der phrasalen Ebene untersuchen wir z. B. Nomen-Präposition- oder Adjektiv-Nomen-Kookkurrenzen wie *Achtung auf...* (cf. Abb. 4) bzw. *eingeschriebenes Kind*. Im Fall von Kookkurrenzen können systematische Auszählungen etwa der mit dem Nomen *Achtung* verwendeten Präpositionen erstellt werden (cf. Abb. 5).

xxx

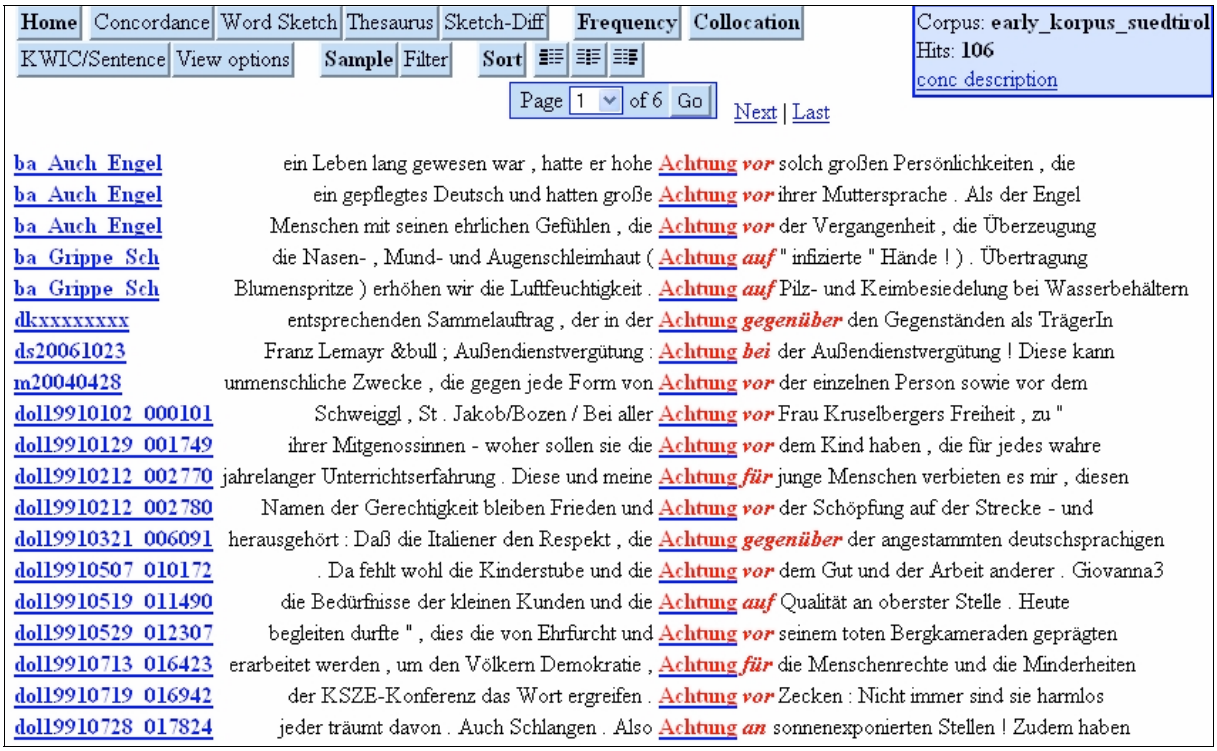


Abb. 4: Anfrageergebnis für *Achtung* + Präposition

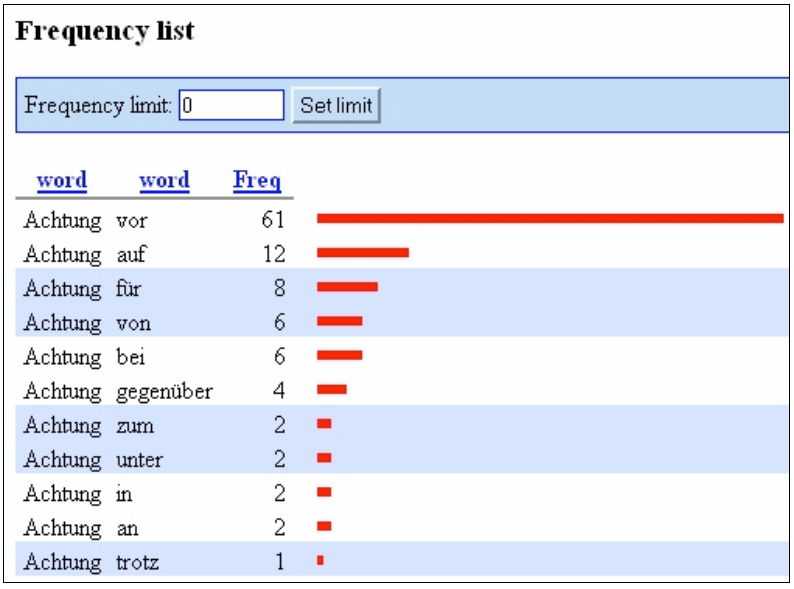


Abb. 5: Häufigkeiten der mit dem Nomen *Achtung* verwendeten Präpositionen

Zur Beantwortung der Frage, wodurch Besonderheiten zu Stande kommen und wie die Ursachen klassifiziert werden können, ist eine manuelle Bearbeitung der Ergebnisdaten notwendig, vor allem auch da Sprache immer durch mehrere Faktoren beeinflusst wird, wobei jeweils nicht ein einzelner Faktor als Ursache benannt werden kann. Einflussgrößen sind neben Umgebungsdialekten auch Kontaktsprachen (im Fall von Südtirol v. a. das Italienische), allgemeine Sprachwandelprozesse und individuelle sprachliche Hintergründe.

Während unverändert übernommene Wörter größtenteils automatisch identifiziert werden können, müssen *Ähnlichkeiten* zum Südtiroler Dialekt, der Kontaktsprache Italienisch (etwa Auswirkungen von direkter Übersetzung) oder auch zu anderen deutschen Varietäten jeweils

von Hand erfasst werden. Als Beispiel für unverändert übernommene Wörter kann *Carabiniere* ('Polizist') genannt werden, für Lehnbildungen z. B. Kondominium (*il condominio* 'das Mehrfamilienhaus').

3.1 Wortebene

Für die systematische Extraktion von lexikalischen und morphologischen Auffälligkeiten auf Wortebene wurden folgende Schritte unternommen:

Aus den vom statistischen TreeTagger (cf. o.) mit Wortarten und Lemmata annotierten Daten des Korpus wurden die Wörter herausgefiltert, die der für das Bundesdeutsche entwickelte TreeTagger auf Grund fehlender Lexikoneinträge nicht lemmatisieren konnte. Neben Tipp- oder Tokenisierungsfehlern wird dabei auch "Spezialwortschatz" mit dem Lemma "unknown" versehen. Ein Auszug aus der resultierenden ca. 5 Millionen Wörter beinhaltenden Liste ist in Tab. 1 gezeigt.

43426	Meran
17540	Durnwalder
14851	Jänner
14557	SVP
9306	Carabinieri
1348	Ortsobmann
584	Verkostung
301	Gastbetrieb
106	Waal
98	Stammrolle
28	Unterfertigte
18	Halbmittag
6	Geschenksidee

Tab. 1: Auszug aus der Liste der "unknown" Lemmata

Diese Liste von "objektiven Besonderheiten" konnte dann reduziert werden durch automatischen Abgleich mit bereits erstellten Sammlungen von speziellem Südtiroler Wortschatz. So wurden z. B. Südtiroler Ortsnamen, Vor- und Nachnamen, Austriazismen und italienische Wörter herausgefiltert. Zusätzlich wurde eine Liste herangezogen, die bereits im Innsbrucker Projekt "Datenbank zum Südtiroler Deutsch"[10] und in Abfalterer (2005) bzw. im Variantenwörterbuch des Deutschen (Ammon et al. 2004) erfasste Südtirolismen enthält, womit südtirolspezifische Wörter ("primäre" als nur in Südtirol gebrauchte und "sekundäre" mit Überschneidungen mit einer anderen Varietät) bezeichnet werden.[11]

Aus einem Abgleich mit solchen Wortsammlungen resultiert nun eine Liste, die u. a. mögliche Ergänzungen für die Wortsammlungen sowie u. U. noch nicht erfassten Südtiroler Spezialwortschatz enthält (cf. Tab. 2).

[10] http://www.uibk.ac.at/projects/woerterbuch/sued/sued.html

[11] Dank gilt H. Abfalterer und allen Südtiroler Ämtern, die Wortlisten zur Verfügung gestellt haben, sowie S. Petrakis und A. Hagelstein für die computerlinguistische Verarbeitung.

43426	Meran	[Ortsname]
17540	Durnwalder	[Nachname]
14851	Jänner	[Austriazismus]
14557	SVP	[Eigenname]
9306	Carabinieri	[Italianismus]
1348	**Ortsobmann**	**⇒ ÜBERPRÜFEN**
584	Verkostung	[Austriazismus]
301	**Gastbetrieb**	**⇒ ÜBERPRÜFEN**
106	Waal	[Südtirolismus]
98	Stammrolle	[Südtirolismus]
28	Unterfertigte	[Austriazismus]
18	Halbmittag	[Südtirolismus]
6	Geschenksidee	[Austriazismus]

Tab. 2: klassifizierte Liste der "unknown" Lemmata

Mit dieser Methode können somit bisherige Sammlungen von Besonderheiten auf Wortebene ggf. ergänzt werden, nicht zuletzt auch im Hinblick auf ein Südtirol-spezifisches Rechtschreibprüfungsprogramm.

3.2 Phrasenebene

In einer weiteren Studie wurden Adjektiv-Nomen-Kookkurrenzen aus dem Dolomiten-Subkorpus extrahiert und mit ebensolchen aus dem Korpus der Frankfurter Rundschau von Juli 1992 bis März 1993 (ca. 40 Millionen Tokens) verglichen. Die Kookkurrenzen, die im Dolomitenkorpus deutlich häufiger auftreten als im Korpus der Frankfurter Rundschau, wurden einer näheren Betrachtung unterzogen; ein Auszug ist in Tab. 3 dargestellt.

[PoS="ADJ"]	[PoS="N"]	*Dolomiten*	*Frankfurter Rundschau*
ganz	Südtirol	1286	0
öffentlich	Hand	1187	247
kommend	Saison	957	259
europäisch	Akademie	592	0
landwirtschaftlich	Grün	464	0
heurig	Saison	318	0
aktiv	Wehrmann	140	1
öffentlich	Grün	103	11
elterlich	Hof	79	1
weiß	**Stimmzettel**	**75**	**2**
regional	Produkt	73	0
bäuerlich	Leben	62	3

Tab. 3: Adjektiv-Nomen-Kookkurrenzen mit Vorkommenshäufigkeiten

Die automatisch erstellten Frequenzvergleichslisten werden nun manuell sortiert, wobei drei grobe Kategorien unterschieden werden. In die erste Klasse fallen linguistisch nicht auffällige Kookkurrenzen, die schlicht durch tatsächliche Gegebenheiten in Südtirol begründet sind (z. B. *Europäische Akademie*). Eine zweite Gruppe bilden Kookkurrenzen, die lexikalische Besonderheiten lediglich in den Kookkurrenzteilen aufweisen, so z. B. das *öffentliche Grün*, was eine öffentliche Grünfläche bezeichnet, oder die *heurige Saison*, die in deutschen Vergleichstexten etwa als *laufende Saison* vorkommt. Die dritte und für unsere Zwecke aufschlussreichste Klasse enthält nun phrasale Auffälligkeiten, die auf Stellen im sprachlichen System hinweisen können, die unter Umständen besonders fehleranfällig sind. So ist der *weiße Stimmzettel* ein direkt aus dem Italienischen übersetzter Begriff (*la scheda bianca*), wobei ein 'nicht ausgefüllter, ungültiger Stimmzettel' gemeint ist.

Listen von möglichen Besonderheiten des Südtiroler Deutschen (mit Ursache z. B. im Dialekt) können also wie beschrieben semi-automatisch aus Korpora extrahiert werden, indem auf bundesdeutschen Daten begründete Lemmatisier-Ergebnisse untersucht oder Vorkommensfrequenzen von Kookkurrenzen in Varietätenkorpora verglichen werden. Auf Wortebene wird dabei z. B. *Ortsobmann* herausgefiltert, auf phrasaler Ebene etwa eine Kookkurrenz wie *weißer Stimmzettel*.

Komplexere Extraktionsmethoden sind notwendig für lexikalische Einheiten, die in Südtirol lediglich eine abweichende oder eine zusätzliche Bedeutung haben, wie z. B. *Stundenplan*, der die Öffnungszeiten einer Tankstelle bezeichnen kann oder *ausrasten*, das verwendet wird, um 'sich ausruhen' auszudrücken. Solche Ausdrücke kommen auch in Vergleichskorpora vor und eine automatische Unterscheidung der Lesarten kann nur mit ausgefeilten computerlinguistischen Semantik-Verarbeitungswerkzeugen wie z. B. der Kontextvektor-methode durchgeführt werden.

Zur Klassifikation der Herkunft und Ursache der einzelnen Abweichungen, um z. B. den Einfluss des Umgebungsdialekts auf eine Standardsprache zu untersuchen, ist nach wie vor eine manuelle Sichtung notwendig; die Grundlage dafür ist jedoch bereits automatisch vorsortiert und auf noch nicht erfasste bzw. wahrscheinliche Besonderheiten reduziert sowie mit quantitativen Daten versehen.

In einer weiterzuführenden semi-automatischen Untersuchung können so systematische Unterschiede des Südtiroler Deutschen zu anderen Varietäten festgestellt und aufgeführt werden, nicht zuletzt auch um der Frage nach Standard und Norm für Südtirol weiter nachzugehen und damit Deutschlernern in Südtirol den Zweit- bzw. Fremdsprachenerwerb zu erleichtern.

4 Ausblick

Im Rahmen von längerfristigen Folgeprojekten werden u. a. die drei folgenden Schwerpunkte gesetzt: (i) die Vergrößerung des thematisch und zeitlich ausgeglichenen Korpus, (ii) dessen Verwendung für weitere systematische quantitative und qualitative linguistische Studien und (iii) die Nutzung deren Ergebnisse zur Unterstützung der Sprachdidaktik (Aston 2001, Ludewig 2005, Zanin 2007).

Um das Korpus Südtirol für aussagekräftige Studien mit statistischer Auswertung verwenden zu können, werden Textdaten hinzugefügt, wobei auch eine bessere Ausgeglichenheit und Repräsentativität des Korpus erreicht werden soll. Dazu werden Texte verschiedener Dekaden (synchron wie diachron), Domänen und Textsorten eingeschlossen, so z. B. spezielle Fachbücher, Gebrauchstexte, übersetzte parallele Texte oder auch Lernertexte. In systematischen Studien erfasste quantitativ untermauerte Unterschiede werden dann als Grundlage dafür dienen, für Editoren und Lerner des Deutschen in Südtirol angepasste Arbeitsmaterialien zu entwickeln, die diese Besonderheiten herausarbeiten und das Bewusstsein dafür schärfen.

Literaturangaben

Abfalterer, Heidemaria (2005): *Der Südtiroler Sonderwortschatz aus plurizentrischer Sicht. Lexikalisch-semantische Besonderheiten im Standarddeutsch Südtirol*. Dissertationsschrift, Universität Innsbruck.

Ammon, Ulrich (1995): *Die deutsche Sprache in Deutschland, Österreich und der Schweiz. Das Problem der nationalen Varietäten*. Berlin/New York.

Ammon, Ulrich et al. (2004): *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol*. Berlin.

Aston, Guy (ed.) (2001): *Learning with Corpora*. Bologna.

Evert, Stefan (2005): *The CQP query language tutorial*. Technical report, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart. http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench.

Kilgarriff, Adam et al. (2004): *The Sketch Engine*. Proceedings Euralex, Lorient. http://www.sketchengine.co.uk, Stand 13.12.2007.

Lanthaler, Franz/Saxalber, Annemarie (1995): "Die deutsche Standardsprache in Südtirol". In: Muhr, Rudolf/Schrodt, Richard/Wiesinger, Peter (eds.): *Österreichisches Deutsch. Linguistische, sozialpsychologische und sprachliche Aspekte einer nationalen Variante des Deutschen*. Wien: 289–305.

Ludewig, Petra (2005): *Korpusbasiertes Kollokationslernen*. Frankfurt am Main etc. (= *Sprache, Sprechen und Computer* 9).

Riedmann, Gerhard (1972): *Die Besonderheiten der deutschen Sprache in Südtirol*. (= *Duden Beiträge, Sonderreihe. Die Besonderheiten der deutschen Schriftsprache im Ausland* 39). Mannheim.

Riehl, Claudia Maria (1994): "Das Problem von 'Standard' und 'Norm' am Beispiel der deutschsprachigen Minderheit in Südtirol". In: Helfrich, Uta / Riehl, Claudia Maria (eds.): *Mehrsprachigkeit in Europa. Hindernis oder Chance?* Wilhelmsfeld: 149–164. (= *Pro Lingua* 14).

Rizzo-Bauer, Hildegard (1962): *Die Besonderheiten der deutschen Schriftsprache in Österreich und Südtirol*. Mannheim. (= *Duden Beiträge, Sonderreihe. Die Besonderheiten der deutschen Schriftsprache im Ausland* 5).

Saxalber Tetter, Annemarie (1989b): "Dialekt in der Schule. Ein Problem für das diglossische und bilinguale Südtirol". In: Koller, Erwin (ed.): *Bayrisch-österreichische Dialektforschung*. Würzburger Tagung. Würzburg: 394–407. (= *Würzburger Beiträge zur deutschen Philologie*. Bd.1).

Schmid, Helmut (1994): *Probabilistic Part-of-Speech Tagging Using Decision Trees*. Proceedings of International Conference on New Methods in Language Processing. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger, Stand 13.12.2007.

Schwienbacher, Brunhild (1997): *Über den Ultner Dialekt. Struktur und Aufbau einer Tiroler Mundart*. Dissertationsschrift, Universität Padua.

Zanin, Renata (2007): "Korpusinstrumente für Deutsch als Zweitsprache". *Theorie und Praxis. Österreichische Beiträge zu Deutsch als Fremdsprache* 10, 2006.